Postprint version of paper published in Chemometr. Intell. Lab. Syst., 148:20–31 (2015).
The content is identical to the published paper, but without the final typesetting by the publisher.

An Extensive Reference Dataset for Fault Detection and Identification in Batch Processes

Jan Van Impe[a], Geert Gins[a,*]

[a] *KU Leuven, Department of Chemical Engineering*
Chemical & Biochemical Process Technology & Control (BioTeC+)
Gebroeders De Smetstraat 1, 9000 Gent (Belgium).

Abstract

Close process monitoring (i.e., detection and identification of disturbances) is important to achieve high process efficiency and safety. The Tennessee Eastman process is an extensive benchmark dataset for fault detection and identification, but it is only representative for continuous processes because it does not contain the inherent non-stationarity that complicates monitoring of batch processes. Nevertheless, batch processes also play an important role in many types of industry. This paper therefore presents an extensive reference dataset for benchmarking data-driven methodologies for fault detection and identification in batch processes.
The original Pensim model (Birol et al., 2002a) is expanded with sensor noise. By changing the properties of the initial conditions and/or model parameters, four subsets of different complexity are generated, each containing 400 batches with normal operation. To correctly assess the fault detection and identification in batch processes, 15 faults are simulated with various amplitudes and onset times for a total of 22,200 faulty batches for each subset, or 90,400 batches in total.
Analysis of the data indicates that the presented types of process faults and their various amplitudes in each of the four subsets present a suitable benchmark for fault detection and identification in batch processes. The dataset is freely available at *http://cit.kuleuven.be/biotec/batchbenchmark*.

Keywords: Batch processes, Statistical Process Control (SPM), Fault detection, Fault identification, Benchmark dataset

1. Introduction

Modern process industry sees a major push towards safe, sustainable, and more profitable operation. Timely detection and diagnosis of process faults, before they have the opportunity to influence process safety and/or product quality, are of utmost importance to maintain safe operation and reduce or even avoid productivity losses (Venkatasubramanian et al., 2003b; Qin, 2012; Ge et al., 2013). Therefore, considerable research attention has been paid to the area of process monitoring (also called Fault Detection and Identification/isolation; FDI) over the last few decades (Qin, 2012; Ge et al., 2013; Ding, 2014).

The existing process monitoring approaches can be categorized as either model-based or data-driven (Yoon and MacGregor, 2000; Ge et al., 2013).

A model-based monitoring scheme employs available first-principles models of the process under study (such as laws of motion, mass balances, energy balances, known reaction schemes, ...) to detect deviations from normal operation. One of the drawbacks of model-based process monitoring is that it is limited to well-known systems of limited size (Yoon and MacGregor, 2000). Typically, first-principles models are available for mechanical or electrical systems. Chemical, biochemical, steel, pulp and paper, or semiconductor processes contain too much uncertainty (e.g., imperfect mixing, biological variability, ...) or are of a too large scale to build accurate-enough first-principles models in an acceptable time (Yoon and MacGregor, 2000; Venkatasubramanian et al., 2003b; Yao and Gao, 2009; Ge et al., 2013).

Data-driven process monitoring, on the other hand, uses only available process measurements to characterize

*Corresponding author; Fax: +32-16-322.991
Email addresses: jan.vanimpe@cit.kuleuven.be (Jan Van Impe), geert.gins@cit.kuleuven.be (Geert Gins)

the nominal process operation. Next, Statistical Process Monitoring (SPM) is used to detect deviations from this normal situation. A detailed overview of active research directions and successful applications of SPM can be found in, i.a., Venkatasubramanian et al. (2003a), Kourti (2005, 2006), Hwang and Kim (2010), Bogomolov (2011), MacGregor and Cinar (2012), Qin (2012), Aldrich and Auret (2013), Ge et al. (2013), and Ding (2014).

SPM algorithms were originally developed for continuous processes because these processes operate around a steady state regime. Batch processes, on the other hand, present a much greater challenge for monitoring owing to their inherent non-stationarity, finite duration, non-linear response, and batch-to-batch variability (Dahl et al., 1999; Smilde, 2001; Eriksson et al., 2013). Furthermore, batch processes commonly suffer from a lack of suitable in-line instrumentation in practice (Dahl et al., 1999). As a result, most novel techniques for fault detection and identification are still developed almost exclusively for continuous processes. Nevertheless, batch processes are widely used in a broad range of sectors, such as the chemical, pharmaceutical, or life sciences industries (Eriksson et al., 2013). Therefore, the development of proper monitoring tools for batch processes is important (Venkatasubramanian et al., 2003a). In their review of SPM for batch processes, Yao and Gao (2009) and Qin (2012) reach the conclusion that more research is needed before advanced SPM methods (such as those capable of dealing with inherent nonlinearities of batch processes) can be applied in practice.

To properly assess the performance of various fault detection and identification methodologies, reliable and extended benchmarks are needed. For continuous processes, the Tennessee Eastman process published by Downs and Vogel (1993) is widely used to benchmark various control and monitoring strategies (Yin et al., 2012; de Lázaro et al., 2015). Chiang et al. (2001) published an extended reference set for fault detection and identification containing normal operation data and data from 22 different types of process upsets, available at *http://web.mit.edu/braatzgroup/TE_process.zip*. The relevance of a proper, extended benchmark is attested by the 157 citations of Downs and Vogel (1993) indexed on Scopus in the period January 2014–May 2015 (17 months). Of these, 124 papers directly concern process monitoring.

When investigating the most important SPM techniques for batch processes as reviewed by Venkatasubramanian et al. (2003a); Kourti (2005); MacGregor and Cinar (2012); Qin (2012); Aldrich and Auret (2013), and Ge et al. (2013), no benchmark comparable to the Tennessee Eastman process exists for batch processes, either in complexity (number of upsets) or frequency of use. Instead, most authors employ one or more small datasets.

For example, Nomikos and MacGregor (1994, 1995a) used a set of 51 normal and 2 faulty batches of a styrene-butadiene rubber (SBR) polymerization reaction generated with the model of Broadhead et al. (1985) for their initial development of Multi-way Principal Component Analysis (MPCA) and Multiway Partial Least Squares (MPLS) for batch process monitoring. In Nomikos and MacGregor (1995b), they employed a set of 55 industrial two-stage polymerization batches provided by DuPont, of which 8 exhibit bad quality. The same DuPont dataset was used by Rännar et al. (1998) to develop hierarchical PCA monitoring. Wold et al. (1998) use data from an industrial fermentation to develop their alternative MPCA approach. Dahl et al. (1999) employ data from 39 batch runs of an autoclave polymerization.

In their presentation of Batch Dynamic PCA (BDPCA) and Batch Dynamic PLS (BDPLS), Chen and Liu (2002) used the SBR and DuPont datasets in addition to a set of 50 normal and 1 faulty batch of the CSTR problem originally presented by Luyben (1990). Choi et al. (2008) also used the SBR dataset and a simulated batch MMA polymerization (Achilias and Kiparissides, 1992) of 100 normal and 3 faulty batches in the development of their autoregressive PCA (ARPCA) approach.

The SBR and DuPont datasets are also used in the review of van Sprang et al. (2002) and the comparison between global, evolving, and local PCA models for monitoring by Ramaker et al. (2005). These two papers also included three additional datasets: (i) an industrial multi-stage polymerization set of 47 normal and 3 abnormal batches (Kosanovich et al., 1996, again provided by DuPont), (ii) a collection of 67 normal and 3 faulty runs of an industrial batch polymerization of PVC (Tates et al., 1999), and (iii) a biochemical conversion set of 27 normal batches and 1 faulty batch (Bijlsma et al., 1998). Ramaker et al. (2005) also employed 24 normal and 2 faulty batch runs of a fat hardening process originally presented by Smilde and Kiers (1999) as a sixth dataset.

Lee et al. (2004) generated 51 normal and 3 faulty batches using the Pensim simulated penicillin fermentation process of Birol et al. (2002a) to demonstrate SPM via Kernel PCA (KPCA). Jia et al. (2010) used two datasets for Batch Dynamic KPCA (BDKPCA): a toy dataset (50 normal batches, 2 faulty) and Pensim (45

Journal homepage: http://www.journals.elsevier.com/chemometrics-and-intelligent-laboratory-systems/
Original file available at: http://www.sciencedirect.com/science/article/pii/S0169743915002105

normal batches, 2 faulty).

The 2-dimensional DPCA (2D-DPCA) was developed by Lu et al. (2005), Yao and Gao (2007, 2008), and Yao et al. (2009) using a toy problem, but the extensions towards Gaussian Mixture Model 2D-DPCA (GMM-2D-DPCA; Yao et al., 2010), and 2-dimensional DKPCA and 2-dimensional Kernel Hebbian Algorithm (2D-KPCA and 2D-KHA; Zhang et al., 2010) are also tested on Pensim data of, respectively, 50 normal and 50 faulty batches, and 5 normal and 5 faulty batches.

Chen and Chen (2006) used the Pensim (50 normal batches, 1 faulty) and SBR datasets to introduce Multi-Hidden Markov Tree-based MPCA (MHMT-MPCA) monitoring of batch processes. Zhao et al. (2007a) test Generalized Moving Window PCA (GMWPCA) via Pensim (20 normal batches) and an injection molding process (40 normal batches). Kulkarni et al. (2004) combined PCA with Generalized Regression Neural Networks (PCA-GRNN), employing 48 normal and 4 faulty runs of the protein synthesis of Lim et al. (1977) and 50 normal and 8 faulty batches of the penicillin production process of Lim et al. (1986).

Recently, Multi-Scale PCA (MSPCA) for batch processes was proposed by Alawi et al. (2015) and tested on 40 normal and 3 faulty Pensim batches.

Zhao and Shao (2006), Zhang et al. (2007), and Yu (2011) all employed 100 normal and 3 faulty Pensim batches for their presentation of batch monitoring using, respectively Multiway Fischer Discriminant Analysis (MFDA), Kernel FDA (KFDA), and Multiway Kernel Localized FDA (MKLFDA). Yan et al. (2014) proposed Semi-supervised Mixture Discriminant Monitoring (SMDM) as an improvement on MKLFDA using data form an injection molding process.

Lee et al. (2003) and Yoo et al. (2004) respectively generated 50 normal and 1 faulty, and 60 normal and 2 faulty Pensim batches to test SPM via Multi-way Independent Component Analysis (MICA). Albazzaz and Wang (2004) conducted a more extensive test of MICA using Pensim (15 normal, 2 faulty) and DuPont datasets, and a third set of 40 normal runs and 1 faulty run of a simulated semi-batch production of polyol lubricant (Yuan and Wang, 2001). They later employ the same set of 15 normal and 2 faulty Pensim batches and the SBR dataset for Dynamic ICA (DICA) for batch monitoring (Albazzaz and Wang, 2007). Pensim was also used to generate 31 normal and 4 faulty batches for benchmarking Kernel ICA (KICA) by Tian et al. (2009). Ge and Song (2008a) developed a combined multilevel ICA-PCA methodology using the DuPont dataset. Zhao et al. (2009) introduced combined Kernel ICA-PCA (KICA-PCA) employing data from Pensim (30 normal batches, 3 faulty) and from a three-tank system (18 normal batches, 2 faulty).

Zhao et al. (2007b) tested their dissimilarity measures for batch monitoring on a toy dataset and on 101 normal and 3 faulty Pensim batches. Hu and Yuan (2009) generated 250 normal and 4 faulty Pensim batches for SPM by means of Tensor Locality Preserving Projections (TLPP) and also validated his procedure on 16 industrial batches. Alvarez et al. (2010) used 187 normal and 444 faulty (8 types of faults at different magnitudes) Pensim for batch monitoring in the original measurement space—the largest Pensim dataset encountered by the authors.

An industrial dataset from a semiconductor etch process (Wise et al., 1999) consisting of 107 normal and 20 faulty batches is used in the works of Chen and Zhang (2010) and Ge et al. (2011, 2013) to respectively test Gaussian Mixture Models (GMM) and Support Vector Data Description (SVDD) for batch monitoring.

Fault identification for batch processes—if even discussed—mostly occurs after fault detection via analysis of contribution plots, despite their suffering from *fault smearing*, which possibly leading to incorrect diagnosis (Westerhuis et al., 2000; Van den Kerkhof et al., 2013). A few exceptions exist, such MKLFDA, where fault detection and identification occur simultaneously (Yu, 2011).

Classification models present an alternative approach to fault identification: given a set of known process upsets of various types, the model assigns the most probable cause to a detected new upset.

Cho and Kim (2004, 2005) proposed an FDA-based fault classification using data from a simulated PVC polymerization. Hereto, they generated a set of 44 normal batches, and 3500 faulty batches of 5 types because their approach requires a number of faulty batches for classifier training greater than the dimensionality of the batches (in their case, the number of monitored sensors times the number of time points). Cho (2007) tested a KFDA classifier for fault identification on two datasets: the same PVC polymerization and Pensim (60 faulty batches, 5 types). Li and Cui (2009) also employ 60 faulty Pensim (5 types) in their work on Feature Vector Selection FDA using Nearest Feature Lines (FVS-FDA-NFL). No information is provided by Cho and Kim (2004, 2005), Cho (2007), or Li and Cui (2009) on the type of the employed process upsets, their magnitude, or their onset time.

A total of 150 Pensim batches (50 of each of three types of process upsets) was used by Monroy et al. (2012) to test fault identification via Artificial Neural

Journal homepage: http://www.journals.elsevier.com/chemometrics-and-intelligent-laboratory-systems/
Original file available at: http://www.sciencedirect.com/science/article/pii/S0169743915002105

Networks (ANN) and Support Vector Machines (SVM). Information on the types of upset and their (fixed) magnitude is provided, but not on onset time.

Van den Kerkhof et al. (2012) employed a set of 840 faulty Pensim batches (4 types of upsets, 8 amplitudes, 6 possible onset times) to compare the fault detection and identification of MPCA, ARPCA, and BDPCA. More recently, Gins et al. (2015) and Wuyts et al. (2015) also used Pensim to compare the performance of k Nearest Neighbors (k-NN) and Least Squares SVM (LS-SVM) classifier. They employed 200 normal and 6600 faulty batches (1100 each of 6 types), providing full specifications on the type, magnitude and onset time of the various faults.

From the above overview of the most important batch monitoring (fault detection and identification) techniques, it is observed that various authors use different datasets. Pensim is frequently used, but all authors generate their own set of batches. To properly assess the performance of fault detection and identification for batch processes, the following factors must be taken into account: standard batch-to-batch variability, the type, magnitude and onset time of a process upset, and the influence of measurement noise and process upsets on the process and its control loops. Hence, it is clear that the above-employed datasets are not suited for correctly evaluating fault detection and identification methods for batch processes: in most cases, the fault detection sets contain only a handful of faulty batches, and the sets dedicated to fault identification very often do not even specify the types of upsets.

Therefore, this paper describes an extensive dataset for benchmarking fault detection and identification methodologies for batch processes that is made freely available. It is composed of 4 subsets of 400 normal and 22,200 faulty batches each (15 types of upsets of varying magnitude and onset), for a grand total of 90,400 batches. The remainder of this manuscript is structured as follows. Section 2 provides a description of the process model used to generate the benchmark datasets. Section 3 describes the various case studies, and Section 4 describes the various simulated process faults. Next, a brief example of process monitoring of the benchmark dataset is detailed in Section 5. Section 6 provides details on how to obtain the reference dataset. Conclusions are drawn in Section 7.

2. Benchmark Model

The Pensim benchmark model of Birol et al. (2002a) was chosen as the basis to generate the extended benchmark dataset presented in this paper, owing to its popularity in literature. Because it was validated on a pilot-scale installation, the Pensim model yields more representative data for process monitoring compared to some toy problems typically used for illustrating various SPM approaches. In addition, the presence of multiple batch phases poses an additional difficulty for SPM.

Section 2.1 provides a brief description of the Pensim model. Section 2.2 describes the practical implementation.

2.1. Process Model

Pensim is based on the morphological model of Birol et al. (2002b). It describes the growth of biomass and production of penicillin in a fed-batch reactor. Initially, the fermentation is operated in batch mode at high substrate concentrations to stimulate biomass growth. Once the initial substrate is nearly exhausted, the process switches to fed-batch mode to maintain a low but non-zero substrate concentration. Under these stressful conditions, penicillin is produced by the biomass. During the entire operation, the reactor is stirred and aerated to provide the biomass with oxygen. Temperature and pH are controlled via PID loops. Feed rate, feed temperature, aeration rate, agitator power, and cold and hot water temperatures are controlled in open loop.

Detailed descriptions of the mathematical model, its parameter values, and the process installation are presented in Birol et al. (2002a,b).

2.2. Practical implementation

The original Pensim simulation package (*http://simulator.iit.edu/web/pensim/simul.html*) only includes measurement noise on the dissolved oxygen and CO_2 concentrations. This enables unrealistically tight control of the process around its temperature and pH set points, greatly facilitating SPM. Many authors already recognized this problem and manually added measurement noise to the simulation data (e.g., Albazzaz and Wang (2004, 2007), Lee et al. (2004), Yoo et al. (2004), Chen and Chen (2006), Ge and Song (2008b), Alvarez et al. (2010), Yao et al. (2010), Gins et al. (2012b, 2015), Vanlaer et al. (2012), Alawi et al. (2015)). In addition, Pensim only simulates a limited set of process upsets.

The Pensim model is therefore implemented in Raymond (*http://cit.kuleuven.be/biotec/raymond*; Gins et al., 2014) to enable easy specification of measurement noise and simulation of more types of upsets.

For the simulations in this work, a nominal feed rate of 0.06 L/h is chosen for the fed-batch phase. A batch is terminated after a total of 25 L of substrate have been

Postprint version of paper published in Chemometr. Intell. Lab. Syst., 148:20–31 (2015).
The content is identical to the published paper, but without the final typesetting by the publisher.

Table 1: Overview of available measurements.

Progress variables

	Name	Noise σ	Resolution
1.	Time [h]	—	—

State variables

	Name	Noise σ	Resolution
2.	Fermentation volume [m^3]	0.002	0.0001
3.	Biomass concentration [g/L]	0.5	0.01
4.	Substrate concentration [g/L]	0.01	0.001
5.	Penicillin concentration [g/L]	0.02	0.002
6.	Dissolved oxygen [mg/L]	0.004	0.0001
7.	Dissolved CO_2 [mg/L]	0.12	0.001
8.	Reactor temperature [K]	0.1	0.01
9.	pH [–]	0.02	0.001
10.	Reaction heat [cal]	—	—

Manipulated & other variables

	Name	Noise σ	Resolution
11.	Feed rate [L/h]	1%	10^{-5}
12.	Feed substrate concentration [g/L]	0.01	0.001
13.	Feed temperature [K]	0.1	0.01
14.	Aeration rate [L/h]	1%	0.01
15.	Agitator power [W]	1%	0.01
16.	Water flow rate [L/h]	1%	0.01
17.	Cold water temperature [K]	0.1	0.01
18.	Hot water temperature [K]	0.1	0.01
19.	Hot/cold switch [–]	—	—
20.	Base flow rate [mL/h]	1%	10^{-3}
21.	Acid flow rate [mL/h]	1%	10^{-4}
22.	Cumulative base flow [mL]	$\Delta(n)^*$	10^{-3}
23.	Cumulative acid flow [mL]	$\Delta(n)^*$	10^{-4}

*Absolute measurement errors on the cumulative base and acid flows at time n are approximately $\Delta_{\text{flow}}(n) = 0.01\sqrt{\sum_{\nu=1}^{n} \text{flow}(\nu)^2}$.

added, for a total batch duration of approximately 460 h. Initial conditions are randomly determined to introduce additional batch-to-batch variability (see Section 3 for more details). All sensors are sampled every 0.02 h.

Table 1 contains an overview of the process states, manipulated variables, and other variables resulting from the simulation. Compared to PENSIM, feed substrate concentration, cold water temperature, and hot water temperature are available, as are cumulative acid and/or base addition. Please note that not all of these variables can be readily measured online. For example, biomass, substrate, and penicillin concentrations are usually measured offline only every 8–10 h (Birol et al., 2002a). Similarly, the substrate concentration of the feed will in practice not be monitored closely, if at all. The author's experience indicates that incoming coolant temperature is also not commonly measured in practice.

Table 2 provides an example set of process variables that could be used for online monitoring. It should also be noted that *instantaneous* base and acid flow are typically monitored in practice (as also indicated in Ta-

Journal homepage: http://www.journals.elsevier.com/chemometrics-and-intelligent-laboratory-systems/
Original file available at: http://www.sciencedirect.com/science/article/pii/S0169743915002105

Postprint version of paper published in Chemometr. Intell. Lab. Syst., 148:20–31 (2015).
The content is identical to the published paper, but without the final typesetting by the publisher.

Table 2: Example set of process variables that are measured online.

Variable
Time
Fermentation volume
Dissolved oxygen concentration
Dissolved CO_2 concentration
Reactor temperature
pH
Feed rate
Feed temperature[†]
Agitator power
Cooling water flow rate
Base flow rate[‡]
Acid flow rate[‡]

[†]Excluding the feed temperature presents more difficult case studies (see Section 5).

[‡]Cumulative base and acid flows are typically more informative than instantaneous flows.

Table 3: Properties of open-loop process inputs.

Name	Nominal value	RBS amplit.
Feed rate [L/h]	0.06	0.005
Feed substrate conc. [g/L]	600	—
Feed temperature [K]	296	0.5
Aeration rate [L/h]	8	0.3
Agitator power [W]	30	1
Cold water temp. [K]	290	0.5
Hot water temp. [K]	323	0.5

ble 2), but very often exhibit on/off behavior. In this case, the *cumulative* base/acid flows would be more informative than the instantaneous flows.

As in standard Pensim, small slow oscillations are introduced on the open-loop input variables to represent upstream variability and non-perfect control. These oscillations are modeled as rolling averages over 1000 samples of a Random Binary Series (RBS) with amplitudes given in Table 3.

Gaussian measurement noise is added to the measured variables with standard deviations σ as listed in Table 1. The values are taken from Lipták (2003) when available, or from in-house lab expertise otherwise. In addition, the resolution of the sensors is limited to the values provided in Table 1.

Table 4: Retuned PI parameters.

pH control		
Parameter	**Acid**	**Base**
K_P	$-4.8 \cdot 10^{-4}$	$4.8 \cdot 10^{-4}$
K_I	1.6	1.6

Temperature control		
Parameter	**Hot water**	**Cold water**
K_P	5	−70
K_I	0.8	0.5

To compensate for the presence of noise and limited sensor accuracy, the PID loops of Pensim that control reactor temperature and pH are replaced with re-tuned PI controllers. Their parameters are presented in Table 4.

3. Case Studies

Four different datasets are presented in this work. The nature of each set is different from the others, and different SPM methodologies are expected to yield better results on different datasets.

Set 1 (Section 3.1) represents a base case where all initial conditions are drawn from normal distributions as is typically done in Pensim. Employing uniform and other non-Gaussian distributions for the initial conditions, set 2 (Section 3.2) is more complex. Sets 3 and 4 (Sections 3.3 and 3.4) include batch-to-batch variation of the model parameters in addition to using non-Gaussian initial conditions.

3.1. Dataset 1: "Base"

3.1.1. Specification

The first dataset is a base case where the initial fermenter volume V_0, biomass concentration $C_{x,0}$, and substrate concentration $C_{s,0}$ are all independently sampled from normal distributions $\mathcal{N}(\mu;\sigma)$ with mean μ and standard deviation σ. To avoid outliers in the initial conditions, values are limited to $\mu \pm 2.5\sigma$.

$$V_0 \sim \mathcal{N}(102.5; 5) \in [90, 115]$$
$$C_{x,0} \sim \mathcal{N}(0.125; 0.03) \in [0.05, 0.20]$$
$$C_{s,0} \sim \mathcal{N}(17.5; 1) \in [15, 20]$$

A total of 400 normal batches (without disturbances) are simulated.

Journal homepage: http://www.journals.elsevier.com/chemometrics-and-intelligent-laboratory-systems/
Original file available at: http://www.sciencedirect.com/science/article/pii/S0169743915002105

Postprint version of paper published in Chemometr. Intell. Lab. Syst., 148:20–31 (2015).
The content is identical to the published paper, but without the final typesetting by the publisher.

3.1.2. Discussion

An initial analysis of the normal batches is performed offline (i.e., comparing complete batches by unfolding the data *batch-wise* as proposed by Nomikos and MacGregor (1994)). Figure 1(a) shows the scatter plot of the first two scores of a PCA model with the variables listed in Table 2 as inputs.

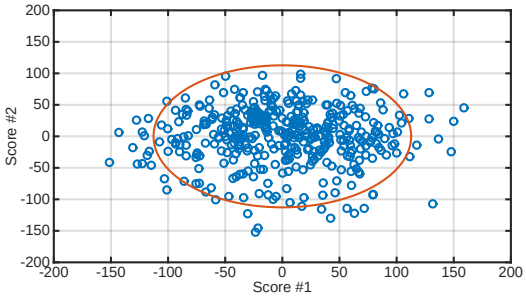
In PCA-based SPM, it is typically assumed that the scores follow a multi-variate normal distribution, and that Hotelling's T^2 statistic—with control limits following an ellipse around the origin of the scores space—is suitable for detecting outliers in the scores space. While the first score approximately follows a normal distribution, the second score shows some deviation from normality, as evidenced by the lack of points falling outside the 95% confidence ellipse near the top of the figure[1]. This suggests that traditional PCA-based analysis might not necessarily be appropriate to characterize the process on a batch level and detect faulty batch runs, or that additional preprocessing of the data is required.

A better quantification of the actual distribution of the scores via, e.g., GMM, Kernel Density Estimation (KDE; Epanechnikov, 1969), or Neighborhood Rank Difference (NRD; Bhattacharyaa et al., 2015) rather than assuming multi-variate normality could improve monitoring performance. However, these techniques typically assume that a high number of data points—in this case: normal batches—is available. If this is not the case, low-density methods are more appropriate (Tang et al., 2002). Another option is to employ different statistics instead of Hotelling's T^2, such as the $\mathcal{D}^2$ distance proposed by He (2007, 2010).

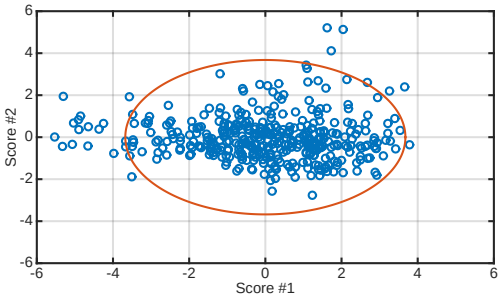
Other approaches for novelty detection, either non-linear extensions such as KPCA, or different methodologies altogether, such as SVDD, One-Class Support Vector Machines (OC-SVM; Schölkopf et al., 2001), One-Class Least-Squares SVM (OC-LS-SVM; Choi, 2009), or novelty detection using k Nearest Neighbors (kNN; He, 2010) offer a third alternative. It should also be investigated whether the decomposition of the measurement profiles in different scales via MSPCA leads to scores that better follow a multi-variate normal distribution.

The online characteristics of the 400 batches are evaluated using MWPCA, where a separate PCA model is constructed for each point in time. Figure 1(b) shows a scatter plot of the first two scores at 5% batch completion with more pronounced violations of the normality assumption. This leads to the same conclusions and suggestions for improving SPM. For online monitoring, however, it is possible that approaches such as BDPCA, ARPCA, PCA with Decorrelated Residuals (PCA-DR; Rato and Reis, 2013) or Sensitivity-Enhancing Transformations (SET; Rato and Reis, 2014a,b) reduce the

[1] It should be noted that the deviation from multivariate normal distribution is quite limited, however. Hence, PCA can still be expected to perform well, just not optimally.



(a)



(b)

Figure 1: Scatter plot of first two PCA scores of (a) offline analysis, and (b) online analysis at 5% completion, for dataset 1 ("base"). The ellipse indicates the 95% confidence bound of Hotelling's T^2.

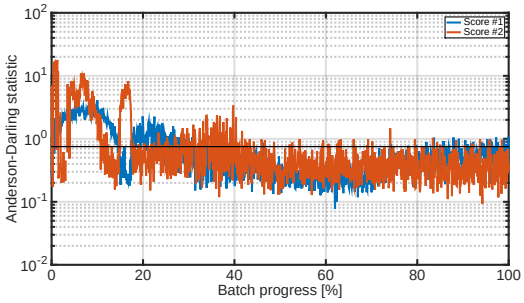


Figure 2: Evolution of the Anderson–Darling statistic for normality testing for the first two scores using MWPCA for dataset 1 ("base"). The horizontal line indicates the minimal value for rejecting the normality assumption.

Journal homepage: http://www.journals.elsevier.com/chemometrics-and-intelligent-laboratory-systems/
Original file available at: http://www.sciencedirect.com/science/article/pii/S0169743915002105

Postprint version of paper published in Chemometr. Intell. Lab. Syst., 148:20–31 (2015).
The content is identical to the published paper, but without the final typesetting by the publisher.

non-normality of the scores and/or residuals by better capturing the dynamics of batch processes.

Figure 2 shows the evolution of the Anderson–Darling statistic for testing normality of the first two scores over the course of the batches. This plot indicates that the multi-variate normality assumption is violated during initial phase of the batch by one of the first two scores, indicating potential sub-optimal monitoring performance by MWPCA. (Scores 3 and 4 follow a normal distribution for almost all time points.) However, investigation of the scores' scatter plots at several time points suggests that MWPCA-based monitoring might still perform well despite some of its assumptions being violated.

3.2. Dataset 2: "Skewed"

3.2.1. Specification

The second dataset introduces non-Gaussian distribution of the initial conditions. The initial volume V_0 is still sampled from the same normal distribution as in the first dataset, independent of the initial biomass and substrate concentrations. The initial biomass concentration $C_{x,0}$ is now uniformly distributed over the range [0.050.20], and the initial substrate concentration $C_{s,0}$ depends nonlinearly on $C_{x,0}$.

$$
\begin{aligned}
V_0 &\sim \mathcal{N}(102.5; 5) \in [90, 115] \\
C_{x,0} &\sim \mathcal{U}(0.05; 0.20) \\
C_{s,0} &\sim \mathcal{N}\left(16 + f_\mu(C_{x,0}); f_\sigma(C_{x,0})\right) \in [15, 20]
\end{aligned}
$$

$$
\begin{aligned}
f_\mu(C_{x,0}) &= \frac{\frac{2.5}{0.09}(C_{x,0} - 0.05)^2 (C_{x,0} - 0.20)}{0.09(C_{x,0} - 0.14) - 0.06(2 \cdot C_{x,0} - 0.19)} \\
f_\sigma(C_{x,0}) &= 1.25 - 4.5 \cdot C_{x,0}
\end{aligned}
$$

Figure 3 displays the relationship between $C_{x,0}$ and $C_{s,0}$. This results in a skewed distribution of $C_{s,0}$, as illustrated in Figure 4. Again, limits are placed on the distributions to avoid outliers. Median and average values of V_0, $C_{x,0}$, and $C_{s,0}$ are close to those employed in the "base" set.

As with the "base" set, 400 normal batches are simulated.

3.2.2. Discussion

When analyzing the second dataset offline with batch-level PCA, deviations from multi-variate normality are observed in the scatter plot depicted in Figure 5(a). The T^2 statistic will fail to detect faulty batches that lie in the upper part of the scores plot but still within the confidence ellipse. In addition, too many normal batches (e.g., some of the batches in towards the side of the scores plot) will be falsely labeled as faulty.

In general, the analysis formulated in Section 3.1 also applies here. Furthermore, because the scores in Figure 5(a) appear to be located in a rectangular region of the scores space, ICA and its variants might be very well suited for this subset.

Figure 6 again gives the evolution of the Anderson–Darling statistic as a function of time. During the first 40% of the batch, the first two scores deviate from normality. This corresponds with a *comet-like* scatter plot as in Figure 5(b) for most time points: a central core of data points with a tail extending to one side. Occasionally, scores plots similar to Figure 5(a) are found.

Visual analysis of the scores plots indicates that the deviations from multi-variate normality are much larger than in the "base" dataset. Hence, more room for improvement on standard PCA-based monitoring exists for the "skewed" dataset.

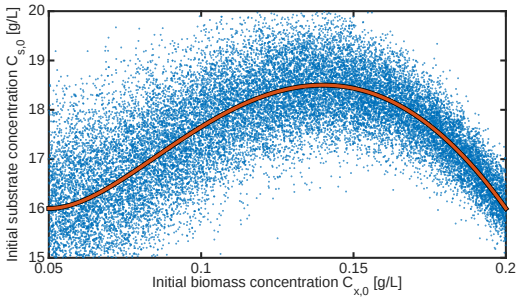


Figure 3: Evolution of initial substrate concentration $C_{s,0}$ as a function of initial biomass concentration $C_{x,0}$. The curve indicates the evolution of the mean $f_\mu(C_{x,0})$. The dependence of f_σ on $C_{x,0}$ is also evident.

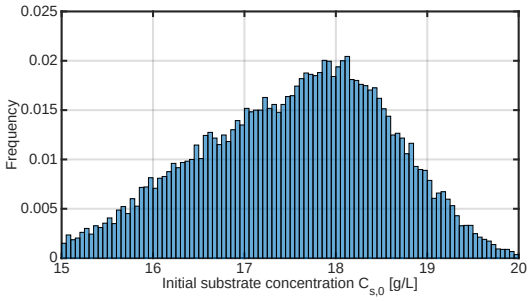


Figure 4: Distribution of the initial substrate concentration $C_{s,0}$ over all simulated batches of dataset 2.

Journal homepage: http://www.journals.elsevier.com/chemometrics-and-intelligent-laboratory-systems/
Original file available at: http://www.sciencedirect.com/science/article/pii/S0169743915002105

Postprint version of paper published in Chemometr. Intell. Lab. Syst., 148:20–31 (2015).
The content is identical to the published paper, but without the final typesetting by the publisher.

3.3. Dataset 3: "Tail"

3.3.1. Specification

In the third dataset, batch-to-batch variability is added on one of the model parameters to further introduce non-Gaussianity in the process. More specifically, the proportionality constant γ in the PENSIM model is drawn from a χ^2 distribution, depicted in Figure 7.

$$\gamma \sim 5 \cdot 10^{-6} + 3.125 \cdot 10^{-6} \cdot \chi^2_2$$

The distribution is chosen so that median and average values of γ are close to the nominal PENSIM value of 10^{-5} mol H^+/g biomass.

Because the paramater γ determines the moles of H^+ ions consumed in the generation of one gram of biomass, variations in γ mainly lead to batch-to-batch variations in pH control actions (acid and base addition). If the control loops saturate, however, pH will deviate more from its set point, influencing biomass growth, substrate consumption, penicillin production, and all other states as secondary factors. This effect is most pronounced towards the end of the batch, where the maximal acid flow is sometimes insufficient to maintain a constant pH for larger values of γ.

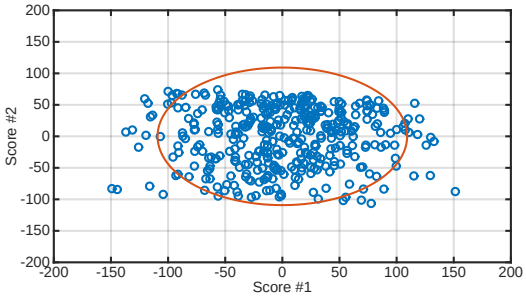
Initial conditions are sampled from the same distributions as used for the "skewed" data.

Again, 400 normal batches are simulated for process characterization.

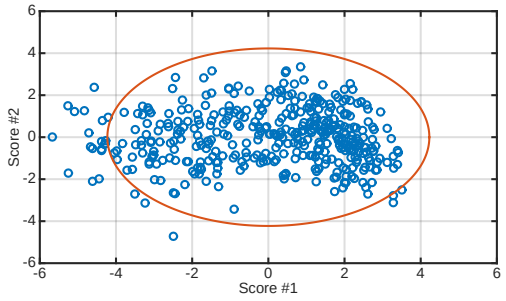
3.3.2. Discussion

Batch-level analysis of the "tail" dataset reveals that the batch-to-batch variation of the parameter γ smooths out the scatter plot in Figure 8(a), which more closely resembles a multi-variate normal distribution than observed for the "skewed" set. The conclusions formulated in Sections 3.1 and 3.2 also apply here.

Moving window analysis again indicates that the normality assumption for the scores is not rejected during most of the batch, as depicted in Figure 9. During the initial 20% of the batch, both scores deviate from normality. Score plots similar to Figure 8(b) are encountered during this time frame. For these plots, which can be best described as a *comet-like* structure between two straight boundaries, the T^2 statistic will result in too



(a)



(b)

Figure 5: Scatter plot of first two PCA scores of (a) offline analysis, and (b) online analysis at 13% completion, for dataset 2 ("skewed"). The ellipse indicates the 95% confidence bound of Hotelling's T^2.

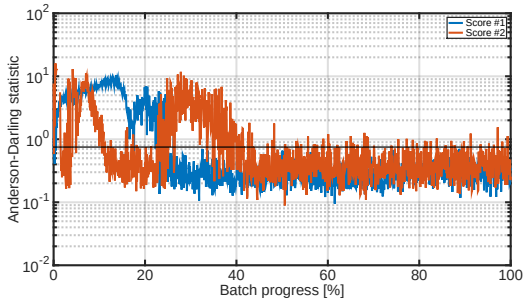


Figure 6: Evolution of the Anderson-Darling statistic for normality testing for the first two scores using a MWPCA for dataset 2 ("skewed"). The horizontal line indicates the minimal value for rejecting the normality assumption.

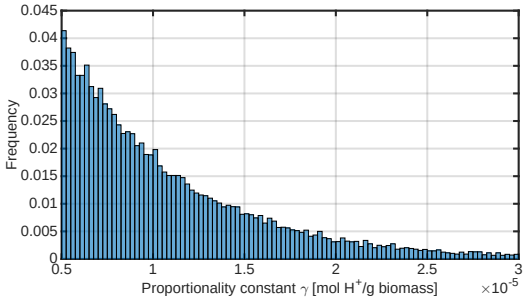
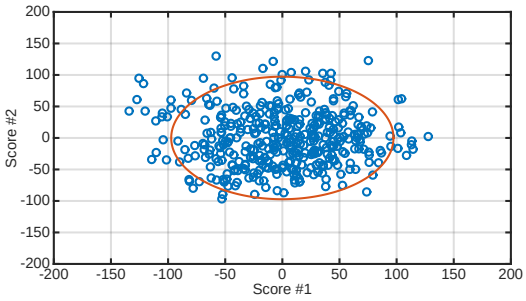


Figure 7: Distribution of the proportionality constant γ over all simulated batches of dataset 3.

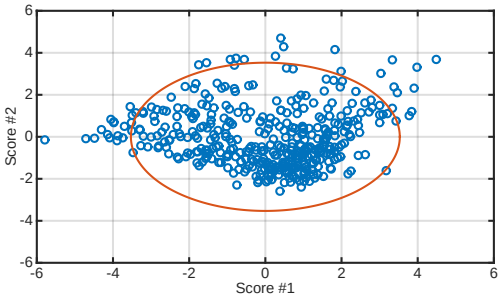
Journal homepage: http://www.journals.elsevier.com/chemometrics-and-intelligent-laboratory-systems/
Original file available at: http://www.sciencedirect.com/science/article/pii/S0169743915002105

Postprint version of paper published in Chemometr. Intell. Lab. Syst., 148:20–31 (2015).
The content is identical to the published paper, but without the final typesetting by the publisher.

many false alarms (towards the left edge and upper right corner of the plot), while at the same leading to late or missed detections (in the lower of the 95% confidence ellipse not populated by any normal batches). During this time frame, score plots similar to Figure 5(a) are also encountered.



(a)



(b)

Figure 8: Scatter plot of first two PCA scores of (a) offline analysis, and (b) online analysis at 13% completion, for dataset 3 ("tail"). The ellipse indicates the 95% confidence bound of Hotelling's T^2.

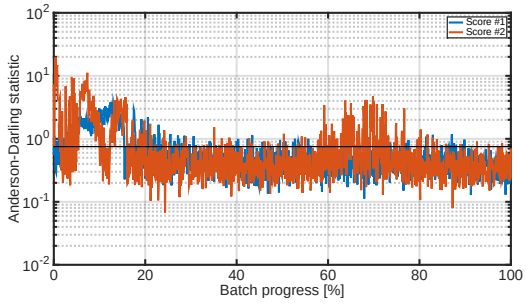


Figure 9: Evolution of the Anderson-Darling statistic for normality testing for the first two scores using a MWPCA for dataset 3 ("tail"). The horizontal line indicates the minimal value for rejecting the normality assumption.

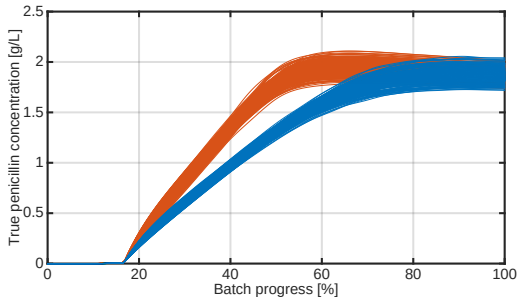
3.4. *Dataset 4: "Two strains"*

3.4.1. *Specification*

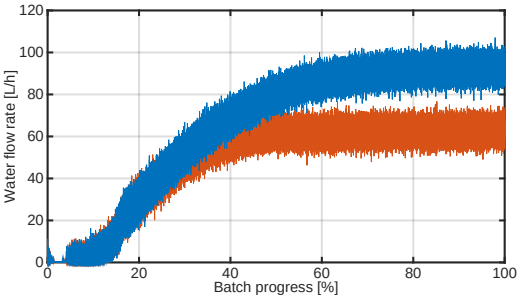
To further increase the inherent complexity of the process, multimodality is introduced in the fourth dataset by employing two types of micro-organisms in the simulation. (This could be interpreted as using strains from two different suppliers, for example.) The two strains differ in yield of product on substrate $Y_{p/s}$ [g penicillin/g substrate], yield of product on oxygen $Y_{p/o}$ [g penicillin/g oxygen] and specific rate of penicillin production μ_p [1/h] as specified in Table 5. The first strain corresponds to the one employed in the "base" (Section 3.1) and "skewed" (Section 3.2) cases. Initial conditions are taken from the same distributions as the "skewed" dataset.

Table 5: Properties of the two biomass strains.

Parameter	Strain 1	Strain 2
$Y_{p/s}$ [g/g]	0.9	0.56
$Y_{p/o}$ [g/g]	0.2	0.124
μ_p [1/h]	0.005	0.0064



(a)



(b)

Figure 10: Evolution of (a) the penicillin concentration [g/L] and (b) the cooling water flow rate for all NOC batches of set 4.

Journal homepage: http://www.journals.elsevier.com/chemometrics-and-intelligent-laboratory-systems/
Original file available at: http://www.sciencedirect.com/science/article/pii/S0169743915002105

Postprint version of paper published in Chemometr. Intell. Lab. Syst., 148:20–31 (2015).
The content is identical to the published paper, but without the final typesetting by the publisher.

Initially, both strains behave similarly. As the batch matures, however, the differences between the two types manifest themselves, as illustrated in Figure 10. While both strains behave differently, Figure 10(a) shows that the final penicillin concentration provides no indication of the existence of two strains. From the example measurement set of Table 2, the existence of two different micro-organism strains can only be observed in the cooling water flow (Figure 10(b)) and—somewhat less obvious—the dissolved oxygen and CO_2 concentrations. The two strains cannot be distinguished from the instantaneous base and acid flow, but cumulative base and acid flows also exhibit bimodal distributions.

As with the other sets, 400 normal batches are simulated, equally distributed over both strains.

3.4.2. Discussion

The presence of two biomass strains is clearly reflected in the batch-wise scores plot. It is evident from Figure 11(a) that standard PCA in combination with the T^2 statistic is not suited for detecting faulty batches. As for the other datasets, more advanced techniques are required to either better characterize the bimodal distribution, replace the T^2 with a different distance measure, or to deal with bimodality directly in the data preprocessing and/or model structure. The various kernel methods (KPCA, KICA, SVDD, OC-(LS)-SVM,...) should be most suited for this dataset, as they can inherently deal with non-convex regions owing to the nonlinear data transformation at their core. GMMs are also suited to deal with this type of data structure.

The bimodal process properties present an additional challenge for batch-end quality prediction as multiple trajectories ultimately result in similar final penicillin concentrations. The accuracy of quality predictions could be improved via the selection of appropriate inputs (e.g. by excluding measurements that display bimodal properties) or by employing non-linear methods. In the former case, however, the prediction model can not be used for fault detection as it will monitor only part of the measurements.

Online analysis of the "two strains" batches via MWPCA results in scores plots similar to the comet-like plots depicted in Figure 5(b) during the initial 20% of the evolution. The normality test in Figure 12 shows deviations from normality for the first score starting from approximately 30% progress. Here, the first score starts capturing the bimodality, as illustrated in Figure 11(b).

3.5. Additional Remarks

Based on the above analysis, it is concluded that the "base" case indeed presents a suitable reference case for monitoring of batch processes because the normal operation data correspond well with the traditional assumption of multi-variate normality. Hence, any difficulties in correct fault detection and isolation will be caused

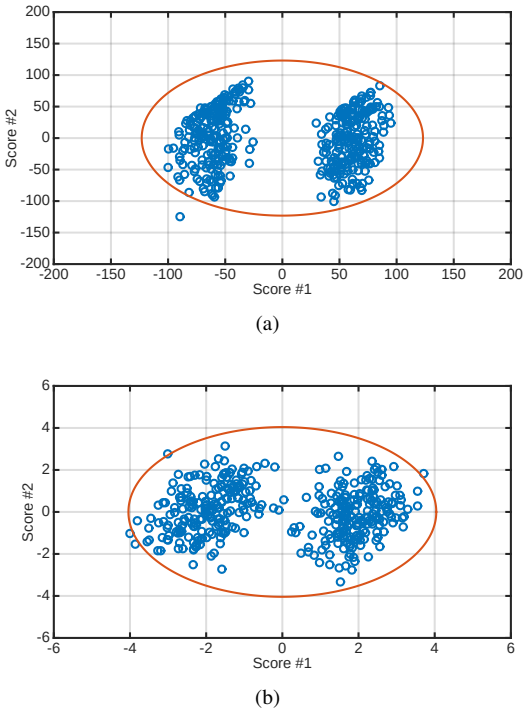


Figure 11: Scatter plot of first two PCA scores of (a) offline analysis, and (b) online analysis at 40% completion, for dataset 4 ("two strains"). The ellipse indicates the 95% confidence bound of Hotelling's T^2.

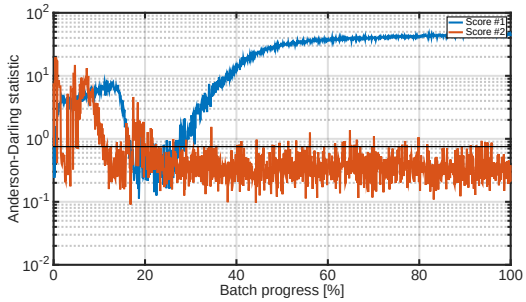


Figure 12: Evolution of the Anderson-Darling statistic for normality testing for the first two scores using a MWPCA for dataset 4 ("two strains"). The horizontal line indicates the minimal value for rejecting the normality assumption.

Journal homepage: http://www.journals.elsevier.com/chemometrics-and-intelligent-laboratory-systems/
Original file available at: http://www.sciencedirect.com/science/article/pii/S0169743915002105

Postprint version of paper published in Chemometr. Intell. Lab. Syst., 148:20–31 (2015).
The content is identical to the published paper, but without the final typesetting by the publisher.

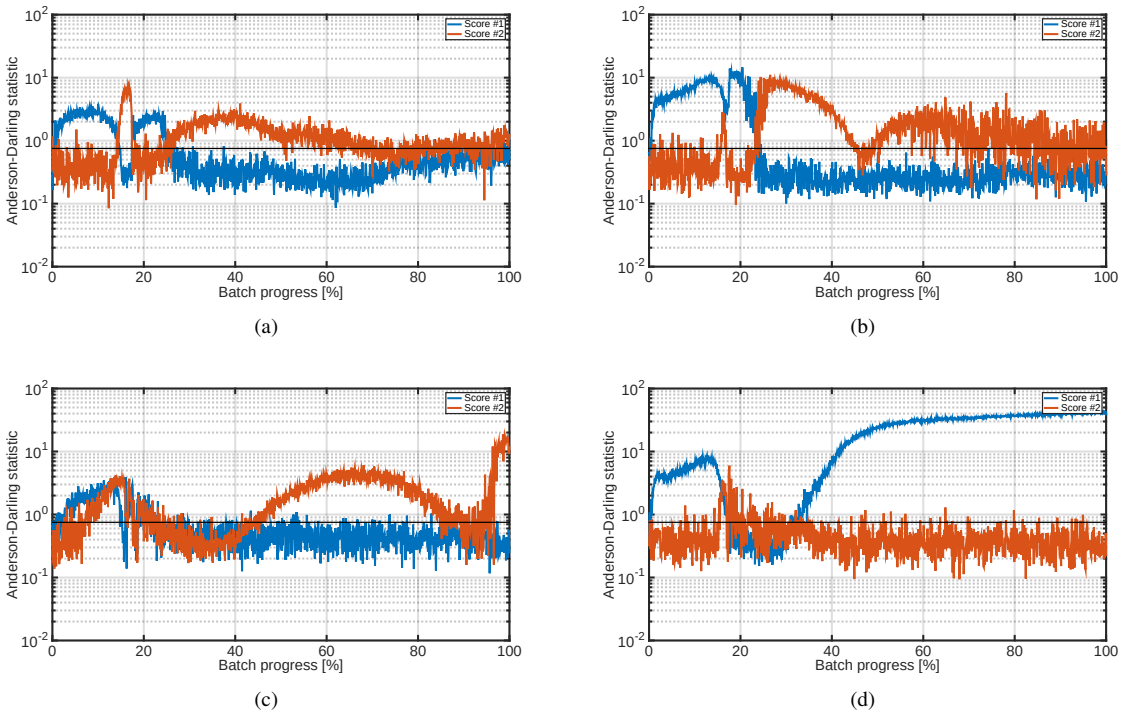


Figure 13: Evolution of the Anderson-Darling statistic for (a) "base", (b) "skewed", (c) "tail", and (d) "two strains" datasets with omitting the cooling water flow rate from the set of monitored variables.

mainly by the properties of the process upsets (type, magnitude, and onset).

The non-normal distributions of the initial conditions in the "skewed" dataset translate into non-normal distributions in the scores space. This presents an extra challenge for SPM owing to the more complicated data structure during normal operation. For the "tail" set, the additional variability on the model parameter γ counteracts some of the effects resulting from non-normal initial conditions. Hence, the "skewed" set likely presents a more difficult challenge for SPM than the "tail" set.

The presence of two micro-organism strains in the "two strains" set leads to multi-modality of the resulting batch data. Therefore, this fourth dataset clearly presents the greatest challenges for fault detection and identification in batch processes compared to the other sets.

It should be stressed that these analyses are valid *only for the given selection of monitored variables and data processing*. For example, when not including the cooling water flow in the monitoring scheme, violations of the normality assumptions are much more frequent as evidenced by Figure 13. In this specific case, the "tail" set exhibits a wider variety of scores plot shapes than the "skewed" set. Therefore, any benchmark study on this dataset should clearly indicate which measurements were included in the SPM model.

4. Process Upsets

For each of the four cases described in Section 3, different process upsets are simulated, as specified in Section 4.1. Section 4.2 provides a discussion of the upsets.

4.1. Types of Upsets

For each of the four cases described in Section 3, 15 different process upsets are simulated, as listed in Table 6. The upsets cover both sudden changes and slow drifts. Some are actual changes in the operation, while others are sensors failures (that possibly impact the fermentation via control loops).

A range of magnitudes is simulated for each upset to present a wide range in detection difficulty for various SPM methodologies. To account for the non-linear, time-varying impact of an upset on the remainder of a batch, the onset time of the upsets is varied. Upsets can

Journal homepage: http://www.journals.elsevier.com/chemometrics-and-intelligent-laboratory-systems/
Original file available at: http://www.sciencedirect.com/science/article/pii/S0169743915002105

Table 6: Overview of simulated proces upsets.

	Fault & Magnitudes
1.	Sudden change in feed substrate concentration −10%, −5%, −2%, −1%, −0.5%, +0.5%, +1%, +2%, +5%, +10%
2.	Change in coolant temperature −2°C, −1°C, −0.5°C, −0.2°C, −0.1°C, +0.1°C, +0.2°C, +0.5°C, +1°C, +2°C
3.	Agitator power drop −0.5%, −1%, −1.5%, −2%, −3%, −4%, −5%, −10%
4.	Aeration rate drop −5%, −10%, −15%, −20%, −25%, −30%, −50%, −70%
5.	Gradual change of feed rate (saturating at 0.04/0.08 L/h for negative/positive drifts) −0.30%/h, −0.15%/h, −0.05%/h, +0.05%/h, +0.15%/h, +0.30%/h
6.	Gradual dissolved oxygen sensor drift (saturating at 0.2/2 for negative/positive drifts) −0.10%/h, −0.05%/h, −0.02%/h, −0.01%/h, −0.005%/h, +0.005%/h, +0.01%/h, +0.02%/h, +0.05%/h, +0.10%/h
7.	Feed temperature change (drift with +1.5°C/h to indicated level) +0.5°C, +1°C, +2°C, +5°C, +10°C, +20°C, +40°C, +60°C
8.	pH sensor drift (saturating at +2) +0.001/h, +0.002/h, +0.003/h, +0.004/h, +0.005/h, +0.010/h, +0.015/h, +0.025/h
9.	Non-functional pH control (no acid or base flow for indicated duration) 0.5 h, 1 h, 2 h, 5 h, 10 h, 20 h
10.	Reduced pH control (control action and maximal control action reduced by indicated fraction) −10%, −20%, −40%, −60%, −80%, −90%
11.	Reactor temperature sensor bias −0.50°C, −0.10°C, −0.05°C, +0.05°C, +0.10°C, +0.50°C
12.	Reactor temperature sensor drift (saturating at −5/+5°C for negative/positive drifts) −0.10°C/h, −0.05°C/h, −0.01°C/h, −0.005°C/h, +0.005°C/h, +0.01°C/h, +0.05°C/h, +0.10°C/h
13.	Reduced temperature control (control action and maximal control action reduced by indicated fraction) −5%, −10%, −20%, −30%, −40%, −50%
14.	Reduced temperature control (control action reduced by indicated fraction, maximal flow not impacted) −10%, −20%, −30%, −40%, −50%, −60%
15.	Contamination (drift of $Y_{P\|S}$ with −0.05/h to indicated level) −0.05, −0.10, −0.15, −0.20, −0.25

start in one of four time ranges: 0–100 h, 100–200 h, 200–300 h, and 300–400 h. A total of 50 repetitions for each combination of onset time interval and fault magnitude ensures statistical representability of monitoring results.[2] This leads to a total of 1000–2000 batches per fault type, or 22,200 faulty batches in total for each of the four case studies.

4.2. Discussion

All of the selected upsets influence the entire fermentation process through the interconnections in the PENSIM model. For example, a change in substrate concentration directly influences the substrate concentration in the fermenter. In turn, this alters the biomass growth rate and the production rate of penicillin, leading to different cooling and pH-control requirements. The associated change in heat generation also leads to changes in evaporative losses and influences the change of fermentation volume. The main differences between the different upsets are the degree to which the various states and control actions are influenced.

The drift of the dissolved oxygen sensor (fault 6) is the only upset that does not impact the operation of the fermentation because aeration and agitation are operated in open loop. This upset is therefore only de-

[2]The "two strains" case includes 25 repetitions of each strain.

tectable from the dissolved oxygen measurements.

Some of the upsets in Table 6 might seem trivial to detect, provided the correct variables are monitored. However, as already noted in Section 2.2, not all process variables are monitored online. In the example set of online measurements of Table 2, it was deliberately chosen not to include the aeration rate in the online measurement set because agitator power drop and aeration rate drop faults influence the process in a similar fashion. As agitator power is measured online, detection and identification of agitator power drops is expected to be straightforward. However, omission of aeration rate from the online measurement set requires this upset to be detected and identified through its propagation to other variables.

Upsets with a large amplitude and early onset present the easiest monitoring task, as the disturbance has enough time to propagate to other measured variables. Small magnitudes in combination with late onset are expected to prove most challenging for SPM.

5. Illustrative Monitoring Results

As a brief illustration, MWPCA was used to monitor the four datasets and detect the presence of process upsets. The models included the measurements listed in Table 2 (using cumulative base and acid flows). For each dataset, 2 principal components were retained for the normal operation PCA model based on a combination of Parallel Analysis (Horn, 1965) and graphical interpretation of the fraction of variation explained by the components. The empirical 95% confidence level was used for the T^2 and Q statistics, which respectively monitor the PCA scores and residuals spaces. More details on the procedure can be found in, e.g. Ramaker et al. (2005).

Figure 14 reports the correct detection rates for the gradual DO sensor drift (fault 6), i.e., the percentage of sample points after the onset of the fault that are detected as such by MWPCA. Each block of 200 batches corresponds to a single drift magnitude from −10%/h for batches 1–200 to +10%/h for batches 1800–2000. Within each block of 200 batches, fault onset time changes from 0–100 h for the first 50 batches to 300–400 h for the final 50 batches.

The authors would like to remark that a fair comparison between monitoring approaches should include information on false alarm information, speed of response, ... in addition to correct detection rates. Moita et al. (2014) proposed a framework hereto, which is currently further developed by Rato et al. (2015).

The detection rates clearly show a more difficult detection for batches with a smaller drift magnitude (center of the graphs) and later onset time (right-hand side of each block of 200 batches). The additional monitoring challenges presented by the "two strains" set is also illustrated in Figure 14. Differences between the "base", "skewed", and "tail" sets are less pronounced, and depend on the specific type of fault.

Fault 7 (change in feed temperature) is found to be the easiest upset to detect, with almost perfect detection rates because the feed temperature is measured directly. To obtain a more challenging study, it is suggested to remove the feed temperature measurements from the set of monitored variables presented in Table 2. In this case, the changing feed temperature must be detected via its propagation to other (measured) variables.

The greatest challenge is posed by faults 1 (change in feed substrate concentration), 9 (non-functional pH control), 10 (reduced pH control), 11 (reactor temperature sensor bias), and 15 (contamination) as low detection rates were obtained for these upsets. The other faults posed a moderate difficulty, with detection rates similar to Figure 14.

It is therefore concluded that the presented process faults are of varied enough type and amplitude to present a good reference for benchmarking various fault detection and identification methodologies.

6. Dataset Description

Each of the four datasets contain 400 normal batches (i.e., without disturbances) for model construction in addition to 22,200 faulty batches, as discussed in Section 4. Sufficient batches of normal operation are available so that some can be kept aside for testing inherent false alarm rates of various SPM approaches and/or tune the control limits of the fault detection statistics. The large diversity in fault types, magnitudes and onset times provides a fair basis for comparing fault detection performance.

The entire dataset of 90,400 batches is made available at *http://cit.kuleuven.be /biotec/batchbenchmark* in aligned and unaligned versions. Each of the four subsets is available in Matlab v7.3 file format, and is approximately 4 GB in size.

In the aligned sets, all batches are brought to equal length via *indicator variables* as employed by Birol et al. (2002a). The volume decrease is used as indicator during the first (batch) phase, and used to resample the full measurement data in this phase every 0.5% decrease for a total of 201 data points. In the second (fed-batch)

Journal homepage: http://www.journals.elsevier.com/chemometrics-and-intelligent-laboratory-systems/
Original file available at: http://www.sciencedirect.com/science/article/pii/S0169743915002105

Postprint version of paper published in Chemometr. Intell. Lab. Syst., 148:20–31 (2015).
The content is identical to the published paper, but without the final typesetting by the publisher.

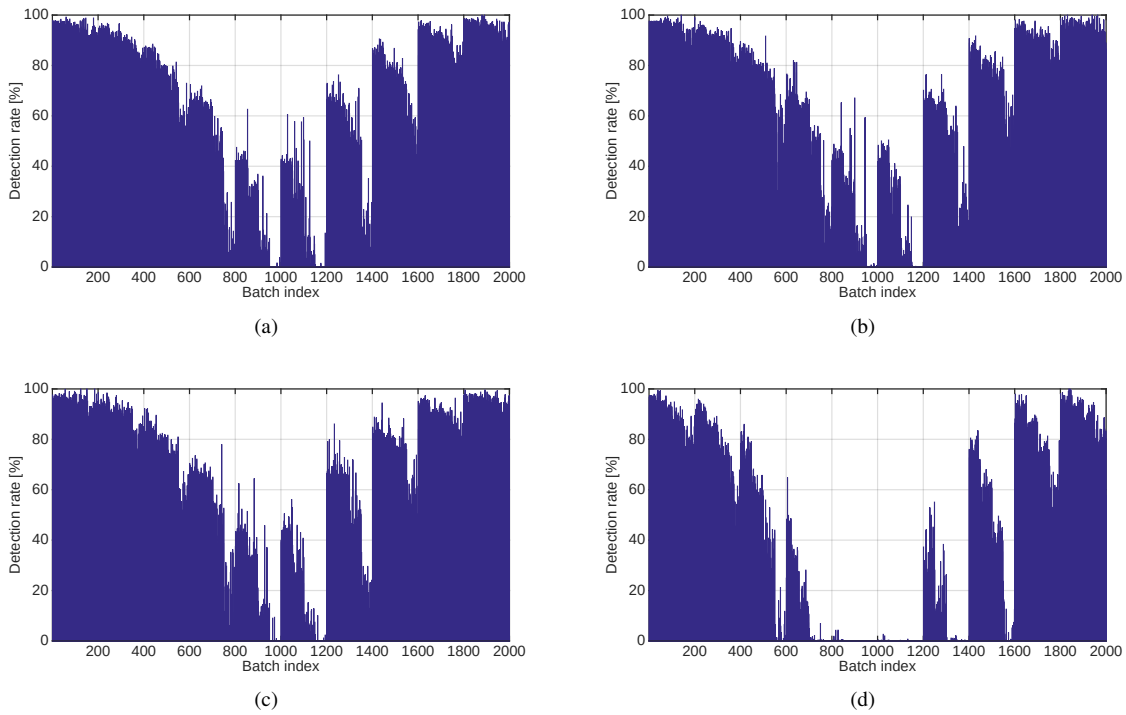


Figure 14: Detection rates for upset 6 (gradual DO sensor drift) for the four cases: (a) "base", (b) "skewed", (c) "tail", (d) "two strains".

phase, the total amount of added substrate is used to resample all batches every 2.5 L added, resulting in 1000 data points in this phase, or a total batch length of 1201 samples.

The unaligned set contains a down-sampled version of the full measurement profiles before synchronization, sampled every 0.2 h rather than every 0.02 h, resulting in approximately 2300 samples per batch. This set of can be used to test monitoring methods that automatically detect the presence of multiple phases (e.g., Zhao et al., 2007a), or to test approaches that do not require batch profile synchronization. In addition, this unaligned set can also be employed to test the effect of profile synchronization on fault detection and identification, albeit in a limited fashion because the batch-to-batch differences in phase duration are small.

Each file contains initial conditions and information on the model parameters (for datasets 3 and 4) for all batches. A total of 24 online measurements are included: the 23 measurements in Table 1, and the exact (unmeasured) penicillin concentration (for quality estimation purposes). For faulty batches, the exact onset time and corresponding sample are reported, as are the exact fault type and its magnitude. Final penicillin concentration is included as final quality variable.

The raw process data (i.e., with sensors sampled every 0.02 h) are also available, albeit upon request only, owing to their total size of 150 GB. The raw data could be used, for example, for testing other data synchronization methods, such as Dynamic Time Warping (DTW; Kassidas et al., 1998), Correlation Optimized Warping (COW; Fransson and Folestad, 2006), and their various variations and combinations (Gins et al., 2012a; Bankó and Abonyi, 2015). Because the growth of biomass and production of penicillin are the major driving forces governing the behavior of the Pensim process, reducing batch-to-batch differences between these (unmeasured) variables could result in better process monitoring.

7. Conclusions

A detailed literature overview of fault detection and identification in batch processes revealed that, while the Tennessee Eastman Process of Downs and Vogel (1993) is available for development and benchmarking of SPM methods for continuous processes, no similar extensive benchmark exists for batch processes. Even though the Pensim model of Birol et al. (2002a) is popular in liter-

Journal homepage: http://www.journals.elsevier.com/chemometrics-and-intelligent-laboratory-systems/
Original file available at: http://www.sciencedirect.com/science/article/pii/S0169743915002105

ature, many authors generate their own dataset. Hence, SPM results cannot be compared directly. In addition, most monitoring approaches for batch processes are evaluated only on a very limited set of faults. This is in stark contrast with the inherent complexity of monitoring batch processes which requires testing process upsets with different magnitudes and onset times to properly assess the performance of various SPM methodologies.

This paper therefore developed an extensive dataset of normal and faulty batch runs based on the PENSIM process, which is made freely available at *http://cit.kuleuven.be/biotec/batchbenchmark* in Matlab v7.3 file format.

After adding representative measurement noise to the various measurements, four subsets of data of different complexity were generated by employing different distributions for the initial conditions of the process and by introducing batch-to-batch variability on some of the model parameters. For each subset, 15 different types of process faults were defined at various fault magnitudes and onset times. This resulted in 400 normal and 22,200 faulty batches for each of the subsets. Analysis of the data demonstrated that the presented dataset forms a good benchmark for developing and testing SPM methods.

Acknowledgements

Work supported in part by Project PFV/10/002 (OPTEC Optimization in Engineering Center) of the Research Council of the KU Leuven, Project KP/09/005 (SCORES4CHEM) of the Industrial Research Council of the KU Leuven, and the Belgian Program on Interuniversity Poles of Attraction IAP VII/19 (DYSCO) initiated by the Belgian Federal Science Policy Office. The authors assume scientific responsibility.

References

Achilias, D., Kiparissides, C., 1992. Development of a general mathematical framework for modeling diffusion controlled free-radical polymerization reactions. Macromolecules 25, 3739–3750.

Alawi, A., Zhang, J., Morris, J., 2015. Multiscale multiblock batch monitoring: Sensor and process drift and degradation. Organic Process Research and Development 19, 145–157.

Albazzaz, H., Wang, X., 2004. Statistical process control charts for batch operations based on independent component analysis. Industrial and Engineering Chemistry Research 43, 6731–6741.

Albazzaz, H., Wang, X., 2007. Introduction of dynamics to an approach for batch process monitoring using independent component analysis. Chemical Engineering Communications 194, 218–233.

Aldrich, C., Auret, L., 2013. Unsupervised process monitoring and fault diagnosis with machine learning methods. Springer-Verlag, London.

Alvarez, C., Brandolin, A., Sanchez, M., 2010. Batch process monitoring in the original measurement's space. Journal of Process Control 20, 716–725.

Bankó, Z., Abonyi, J., 2015. Mixed dissimilarity measure for piecewise linear approximation based time series applications. Expert Systems with Applications 42, 7664–7675.

Bhattacharyaa, G., Ghoshb, K., Chowdhuryc, A., 2015. Outlier detection using neighborhood rank difference. Pattern Recognition Letters 60–61, 24–31.

Bijlsma, S., Louwerse, D., Windig, W., Smilde, A., 1998. Rapid estimation of rate constants of batch processes using on-line SW-NIR. AIChE Journal 44, 2713–2723.

Birol, G., Ündey, C., Cinar, A., 2002a. A modular simulation package for fed-batch fermentation: penicillin production. Computers and Chemical Engineering 26, 1552–1565.

Birol, G., Ündey, C., Parulekar, S., Cinar, A., 2002b. A morphologically structured model for penicillin production. Biotechnology and Bioengineering 77, 538–552.

Bogomolov, A., 2011. Multivariate process trajectories: capture, resolution and analysis. Chemometrics and Intelligent Laboratory Systems 108, 49–63.

Broadhead, T., Hamielec, A., MacGregor, J., 1985. Dynamic modeling of the batch, semi-batch and continuous production of styrene-butadiene copolymers by emulsion polymerization. Makromolekulare Chemie Supplement 10.

Chen, J., Chen, H.H., 2006. On-line batch process monitoring using MHMT-based MPCA. Chemical Engineering Science 61, 3223–3239.

Chen, J., Liu, K.C., 2002. On-line batch process monitoring using dynamic PCA and dynamic PLS models. Chemical Engineering Science 57, 63–75.

Chen, T., Zhang, J., 2010. On-line multivariate statistical monitoring of batch processes using gaussian mixture model. Computers and Chemical Engineering 34, 500–507.

Chiang, L., Braatz, R., Russell, E., 2001. Fault detection and diagnosis in industrial systems. Springer.

Cho, H., 2007. Nonlinear feature extraction and classification of multivariate process data in kernel feature space. Expert Systems with Applications 32, 534–542.

Cho, H., Kim, K., 2004. Fault diagnosis of batch processes using discriminant model. International Journal of Production Research 42, 597–612.

Cho, H., Kim, K., 2005. Diagnosing batch processes with insufficient fault data: generation of pseudo batches. International Journal of Production Research 43, 2997–3009.

Choi, S., Morris, J., Lee, I.B., 2008. Dynamic model-based batch process monitoring. Chemical Engineering Science 63, 622–636.

Choi, Y.S., 2009. Least squares one-class support vector machine. Pattern Recognition Letters 30, 1236–1240.

Dahl, K., Piovoso, M., Kosanovich, K., 1999. Translating third-order data analysis methods to chemical batch processes. Chemometrics and Intelligent Laboratory Systems 46, 161–180.

de Lázaro, J., Moreno, A., Santiago, O., da Silva Neto, A., 2015. Optimizing kernel methods to reduce dimensionality in fault diagnosis of industrial systems. Computers and Chemical Engineering 87, 140–149.

Ding, S., 2014. Data-driven design of monitoring and diagnosis systems for dynamic processes: A review of subspace technique based schemes and some recent results. Journal of Process Control 24, 431–449.

Downs, J., Vogel, E., 1993. A plant-wide industrial process control problem. Computers and Chemical Engineering 17, 245–255.

Epanechnikov, V., 1969. Non-parametric estimation of a multivariate probability density. Theory of Probability and its Applications 14, 153–158.

Eriksson, L., Byrne, T., Johansson, E., Tryg, J., Vikström, C., 2013. Multi- and Megavariate Data Analysis: Basic Principles and Applications. Umetrics Academy, 3rd Revised Edition.

Fransson, M., Folestad, S., 2006. Real-time alignment of batch process data using COW for on-line process monitoring. Chemometrics and Intelligent Laboratory Systems 84, 56–61.

Ge, Z., Gao, F., Song, Z., 2011. Batch process monitoring based on support vector data description method. Journal of Process Control 21, 949–959.

Ge, Z., Song, Z., 2008a. Batch process monitoring based on multilevel ICA-PCA. Journal of Zhejiang University Science A 9, 1061–1069.

Ge, Z., Song, Z., 2008b. Online batch process monitoring based on multi-model ICA-PCA method, in: Proceedings ot the 7th World COngress on Intelligent Controla and Automation, pp. 260–264.

Ge, Z., Song, Z., Gao, F., 2013. Review of recent research on data-based process monitoring. Industrial and Engineering Chemistry Research 52, 3543–3562.

Gins, G., Van den Kerkhof, P., Van Impe, J., 2012a. Hybrid derivative dynamic time warping for online industrial batch-end quality estimation. Industrial and Engineering Chemistry Research 51, 6071–6084.

Gins, G., Van den Kerkhof, P., Vanlaer, J., Van Impe, J., 2015. Improving classification-based diagnosis of batch processes through data selection and appropriate pretreatment. Journal of Process Control 26, 90–101.

Gins, G., Vanlaer, J., Van den Kerkhof, P., Van Impe, J., 2014. The RAYMOND simulation package — Generating RAYpresentative MONitoring Data to design advanced process monitoring and control algorithms. Computers and Chemical Engineering 69, 108–118.

Gins, G., Vanlaer, J., Van Impe, J., 2012b. Discriminating between critical and noncritical disturbances in (bio)chemical batch processes using multimodel fault detection and end-quality prediction. Industrial and Engineering Chemistry Research 51, 12375–12385.

He, Q.P., 2007. Fault detection using the k-nearest neighbor rule for semiconductor manufacturing processes. IEEE Transactions on Semiconductor Manufacturing 20, 345–354.

He, Q.P., 2010. Large-scale semiconductor process fault detection using a fast pattern recognition-based method. IEEE Transactions on Semiconductor Manufacturing 23, 194–200.

Horn, J., 1965. A rationale and test for the number of factors in factor analysis. Psychometrika 30, 179–185.

Hu, K., Yuan, J., 2009. Batch process monitoring with tensor factorization. Journal of Process Control 19, 288–296.

Hwang, I., Kim, S., 2010. A survey of fault detection, isolation, and reconfiguration methods. IEEE Transactions on Control Systems Technology 18, 636–653.

Jia, M., Chu, F., Wang, F., Wang, W., 2010. On-line batch process monitoring using batch dynamic kernel principal component analysis. Chemometrics and Intelligent Laboratory Systems 101, 110–122.

Kassidas, A., MacGregor, J., Taylor, P., 1998. Synchronization of batch trajectories using dynamic time warping. AIChE Journal 44, 864–875.

Kosanovich, K., Dahl, K., Piovoso, M., 1996. Improved process understanding using multiway principal component analysis. Industrial and Engineering Chemistry Research 35, 138–146.

Kourti, T., 2005. Application of latent variable methods to process control and multivariate statistical process control in industry. International Journal of Adaptive Control and Signal Processing 19, 213–246.

Kourti, T., 2006. The process analytical technology initiative and multivariate process analysis, monitoring and control. Analytical and Bioanalytical Chemistry 384, 1043—1048.

Kulkarni, S., Chaudhary, A., Nandi, S., Tambe, S., Kulkarni, B., 2004. Modeling and monitoring of batch processes using principal component analysis (PCA) assisted generalized regression neural networks (GRNN). Biochemical Engineering Journal 18, 193–210.

Lee, J.M., Yoo, C., Lee, I.B., 2003. Online batch process monitoring using different unfolding method and independent component analysis. Journal of Chemical Engineering of Japan 36, 1384–1396.

Lee, J.M., Yoo, C., Lee, I.B., 2004. Fault detection of batch processes using multiway kernel principal component analysis. Computers and Chemical Engineering 28, 1837–1847.

Li, J., Cui, P., 2009. Improved kernel fisher discriminant analysis for fault diagnosis. Expert Systems with Applications 36, 1423–1432.

Lim, H., Chen, B., Creagan, C., 1977. An analysis of extended and exponentially-fed-batch cultures. Biotechnology and Bioengineering 14, 425–433.

Lim, H., Tayeb, Y., Modak, J., Bonte, P., 1986. Computational algorithm for a class of fed-batch fermentation. Biotechnology and Bioengineering 28, 1408–1420.

Lipták, B., 2003. Instrument Engineers' Handbook: Process Measurement and Analysis. volume 1. 4 ed., CRC Press.

Lu, N., Yao, Y., Gao, F., Wang, F., 2005. Two-dimensional dynamic PCA for batch process monitoring. AIChE Journal 51, 3300–3304.

Luyben, W., 1990. Process modeling, simulation and control for chemical engineers. McGraw-Hill, New York.

MacGregor, J., Cinar, A., 2012. Monitoring, fault diagnosis, fault-tolerant control and optimization: Data driven methods. Computers and Chemical Engineering 47, 111–120.

Moita, R., Gomes, V., Saraiva, P., Reis, M., 2014. An extended comparative study of two- and three-way methodologies for the on-line monitoring of batch processes. Computer Aided Chemical Engineering 33, 517–522.

Monroy, I., Villez, K., Graells, M., Venkatasubramanian, V., 2012. Fault diagnosis of a benchmark fermentation process: a comparative study of feature extraction and classification techniques. Bioprocess and Biosystems Engineering 35, 689–704.

Nomikos, P., MacGregor, J., 1994. Monitoring batch processes using multiway principal component analysis. AIChE Journal 40, 1361–1375.

Nomikos, P., MacGregor, J., 1995a. Multi-way partial least squares in monitoring batch processes. Chemometrics and Intelligent Laboratory Systems 30, 97–108.

Nomikos, P., MacGregor, J., 1995b. Multivariate SPC charts for monitoring batch processes. Technometrics 37, 41–59.

Qin, S., 2012. Survey on data-driven industrial process monitoring and diagnosis. Annual Reviews in Control 26, 220–234.

Ramaker, H.J., van Sprang, E., Westerhuys, J., Smilde, A., 2005. Fault detection properties of global, local and time evolving models for batch process monitoring. Journal of Process Control 15, 799–805.

Rännar, S., MacGregor, J., Wold, S., 1998. Adaptive batch monitoring using hierarchical PCA. Chemometrics and Intelligent Laboratory Systems 41, 73–81.

Rato, T., Reis, M., 2013. Fault detection in the Tennessee Eastman benchmark process using dynamic principal components analysis based on decorrelated residuals (DPCA-DR). Chemometrics and Intelligent Laboratory Systems 125, 101–108.

Rato, T., Reis, M., 2014a. Non-causal data-driven monitoring of the process correlation structure: A comparison study with new methods. Computers and Chemical Engineering 71, 307–322.

Rato, T., Reis, M., 2014b. Sensitivity enhancing transformations for monitoring the process correlation structure. Journal of Process Control 24, 905–915.

Rato, T., Rendall, R., Gomes, V., Chin, S.T., Chiang, L., Saraiva, P., Reis, M., 2015. A systematic methodology for comparing batch process monitoring methods: Part I – assessing detection strength.

(In preparation).

Schölkopf, B., Platt, J., Shawe-Taylor, J., Smola, A., Williamson, R., 2001. Estimating the support of a high-dimensional distribution. Neural Computation 13, 1443–1471.

Smilde, A., 2001. Comments on three-way analyses used for batch process data. Journal of Chemometrics 15, 19–27.

Smilde, A., Kiers, H., 1999. Multiway covariates regression models. Journal of Chemometrics 13, 31–48.

Tang, J., Chen, Z., Fu, A., Cheung, D., 2002. Enhancing effectiveness of outlier de- tections for low density patterns, in: Proceedings of the 6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD), pp. 45–84.

Tates, A., Louwerse, D., Smilde, A., Koot, G., Berndt, H., 1999. Monitoring a PVC batch process with multivariate statistical process control charts. Industrial and Engineering Chemistry Research 38, 4768–4776.

Tian, X., Zhang, X., Deng, X., Chen, S., 2009. Multiway kernel independent component analysis based on feature samples for batch process monitoring. Neurocomputing 72, 1584–1596.

Van den Kerkhof, P., Gins, G., Vanlaer, J., Van Impe, J., 2012. Dynamic model-based fault diagnosis for (bio)chemical batch processes. Computers and Chemical Engineering 40, 12–21.

Van den Kerkhof, P., Vanlaer, J., Gins, G., Van Impe, J., 2013. Analysis of smearing-out in contribution plot based fault isolation for Statistical Process Control. Chemical Engineering Science 104, 285–293.

van Sprang, E., Ramaker, H.J., Westerhuys, J., Gurden, S., Smilde, A., 2002. Critical evaluation of approaches for on-line batch process monitoring. Chemical Engineering Science , 3979–3991.

Vanlaer, J., Van den Kerkhof, P., Gins, G., Van Impe, J., 2012. The influence of input and output measurement noise on batch-end quality prediction with partial least squares. Lecture Notes in Computer Science: Advances in Data Mining 7377, 121–135.

Venkatasubramanian, V., Rengaswamy, R., Kavuri, S., Yin, K., 2003a. A review of process fault detection and diagnosis – Part III: Process history based methods. Computers and Chemical Engineering 27, 327–346.

Venkatasubramanian, V., Rengaswamy, R., Yin, K., Kavuri, S., 2003b. A review of process fault detection and diagnosis – Part I: Quantitative model-based methods. Computers and Chemical Engineering 27, 293–311.

Westerhuis, J., Gurden, S., Smilde, A., 2000. Generalized contribution plots in multivariate statistical process monitoring. Chemometrics and Intelligent Laboratory Systems 51, 95–114.

Wise, B., Gallagher, N., Butler, S., White Jr, D., Barna, G., 1999. A comparison of principal component analysis, multiway principal component analysis, trilinear decomposition and parallel factor analysis for fault detection in a semiconductor etch process. Journal of Chemometrics 13, 379–396.

Wold, S., Kettaneh, N., Friden, H., Holmberg, A., 1998. Modelling and diagnostics of batch processes and analogous kinetic experiments. Chemometrics and Intelligent Laboratory Systems 44, 331–340.

Wuyts, S., Gins, G., Van den Kerkhof, P., Van Impe, J., 2015. Fault identification in batch processes using process data or contribution plots: A comparative study. Advanced Control of Chemical Processes 8, 1283–1288.

Yan, Z., Huang, C.C., Yao, Y., 2014. Semi-supervised mixture discriminant monitoring for chemical batch processes. Chemometrics and Intelligent Laboratory Systems 134, 10–22.

Yao, Y., Chen, T., Gao, F., 2010. Multivariate statistical monitoring of two-dimensional dynamic batch processes utilizing non-gaussian information. Journal of Process Control 20, 1188–1197.

Yao, Y., Dao, Y., Lu, N., Lu, J., Gao, F., 2009. Two-dimensional dynamic principal component analysis with autodetermined support region. Industrial and Engineering Chemistry Research 48, 837–843.

Yao, Y., Gao, F., 2007. Batch process monitoring in score space of two-dimensional dynamic principal component analysis (PCA). Industrial and Engineering Chemistry Research 46, 8033–8043.

Yao, Y., Gao, F., 2008. Subspace identification for two-dimensional dynamic batch process statistical monitoring. Chemical Engineering Science 63, 3411–3418.

Yao, Y., Gao, F., 2009. A survey on multistage/multiphase statistical modeling methods for batch processes. Annual Reviews in Control 33.

Yin, S., Ding, S., Haghani, A., Hao, H., Zhang, P., 2012. A comparison study of basic data-driven fault diagnosis and proces monitoring methods on the benchmark Tennessee Eastman process. Journal of Process Control 22, 1567–1581.

Yoo, C., Lee, J.M., Vanrolleghem, P., Lee, I.B., 2004. On-line monitoring of batch processes using multiway independent component analysis. Chemometrics and Intelligent Laboratory Systems 71, 151–163.

Yoon, S., MacGregor, J., 2000. Statistical and causal model-based approaches to fault detection and isolation. AIChE Journal 46, 1813–1824.

Yu, J., 2011. Nonlinear bioprocess monitoring using multiway kernel localized fisher discriminant analysis. Industrial and Engineering Chemistry Research 50, 3390–3402.

Yuan, B., Wang, X., 2001. Multilevel PCA and inductive learning for knowledge extraction from operational data of batch processes. Chemical Engineering Communications 185, 201–221.

Zhang, X., Yan, W., Zhao, X., Shao, H., 2007. Nonlinear biological batch process monitoring and fault identification based on kernel fisher discriminant analysis. Process Biochemistry 42, 1200–1210.

Zhang, Y., Li, Z., Zhou, H., 2010. Statistical analysis and adaptive technique for dynamical process monitoring. Chemical Engineering Research and Design 88, 1381–1392.

Zhao, C., Gao, F., Wang, F., 2009. Nonlinear batch process monitoring using phase-based kernel-independent component analysis-principal component analysis (KICA-PCA). Industrial and Engineering Chemistry Research 48, 9163–9174.

Zhao, C., Wang, F., Gao, F., Lu, N., Jia, M., 2007a. Adaptive monitoring method for batch processes based on phase dissimilarity updating with limited modeling data. Industrial and Engineering Chemistry Research 46, 4943–4953.

Zhao, C., Wang, F., Jia, M., 2007b. Dissimilarity analysis based batch process monitoring using moving windows. AIChE Journal 53, 1267–1277.

Zhao, X., Shao, H., 2006. On-line batch process monitoring and diagnosing based on fisher discriminant analysis. Journal of Shanghai Jiaotong University E-11, 307–312.